

THEORETICAL EVALUATION OF FEATURE SELECTION METHODS BASED ON MUTUAL INFORMATION

CLÁUDIA PASCOAL¹, M. ROSÁRIO OLIVEIRA¹, ANTÓNIO PACHECO¹ AND RUI VALADAS²

¹ *CEMAT and Dep. Mathematics,
Instituto Superior Técnico, Universidade de Lisboa,
Av. Rovisco Pais, 1049-001 Lisboa, Portugal.*

² *IT and Dep. Electrical and Computer Engineering,
Instituto Superior Técnico, Universidade de Lisboa,
Av. Rovisco Pais, 1049-001 Lisboa, Portugal.*

ABSTRACT. Feature selection methods are usually evaluated by wrapping specific classifiers and datasets in the evaluation process, resulting very often in unfair comparisons between methods. In this work, we develop a theoretical framework that allows obtaining the true feature ordering of two-dimensional sequential forward feature selection methods based on mutual information, which is independent of entropy or mutual information estimation methods, classifiers, or datasets, and leads to an undoubtful comparison of the methods. Moreover, the theoretical framework unveils problems intrinsic to some methods that are otherwise difficult to detect, namely inconsistencies in the construction of the objective function used to select the candidate features, due to various types of indeterminations and to the possibility of the entropy of continuous random variables taking null and negative values.

1. INTRODUCTION

Feature selection is an important step in statistical learning problems involving high-dimensional data, e.g. regression and classification. Its main benefits are to facilitate data visualization and data understanding, reduce measurement and storage requirements, reduce training and utilization times, and improve predictor performance [1]. Due to its importance, many excellent surveys on feature selection methods have been produced over the years [2–11].

The main goal of feature selection is to find a subset of features that leads to optimal performance of the learning process. This involves keeping relevant features, and removing those that are irrelevant or redundant [8, 11–15].

Feature selection methods can be classified into three categories [6, 9]: *filters* methods, *wrappers* methods, and *embedded* methods. Wrapper methods embrace the classifier in the selection process; the features are selected according to classifier performance metrics such as recall and precision [6, 16, 17]. Filter methods select the features independently of the classifier; the selection process tries to find the subset of features that is most associated with the class variable [6, 8, 9]. Embedded methods combine the filter selection stage with the learning step [6, 8, 18].

Wrapper methods have two relevant disadvantages: their large computational complexity and their dependence on a specific classifier. Apart from being classifier independent, filter methods are computationally less demanding than wrapper methods and, as a result, are more suitable for high-dimensional problems. Embedded methods are also classifier dependent, but less onerous in computational complexity and less

Key words and phrases. Feature selection, Mutual information, Entropy.

sensitive to over-fitting than wrapper methods. However, embedded methods are designed specifically to a certain classifier, which constrains their generalization [6, 9].

Optimal feature selection is usually unfeasible because the search space grows exponentially with the number of features. As a result various sub-optimal algorithms have been devised, with sequential forward selection being the most commonly adopted solution. Forward selection algorithms start from an empty set of features and add, in each step, the feature that jointly, i.e. together with already selected features, achieves the maximum association with the class (also called maximum relevance). Various approaches have been followed regarding how this association is accounted for.

A widely accepted association measure used in filter methods is Mutual Information (MI) [19], an information-theoretic metric able to capture both linear and non-linear dependencies among random variables. One approach is to estimate directly the high-dimensional MI between the class, the already selected features, and the candidate one. However, this may not be an easy task as, except for low dimensions, the estimation cannot rely on histograms, because of the sparse data distributions often encountered in high-dimensional spaces.

One alternative to the estimation of high-dimensional MI or entropy measures is to use two-dimensional approximations. The usual approach is to rely on a criterion that balances the relevance of a candidate feature with its redundancy to already-selected features. The relevance component is accounted through the MI between the class variable and the candidate feature. The redundancy component involves calculating the MI between the class, the already-selected features, and the candidate feature. This is still a high-dimensional problem, but several approximations were considered to reduce it to two-dimensions [20–24].

There has been an increasing concern around the evaluation of feature selection methods. The common practice is to perform the evaluation considering specific classifiers and datasets. This may explain why there are so many proposals and so little consensus on the best features to be used in particular scenarios. Filter methods are per-definition independent from the classifier and, therefore, should be evaluated independently from the classifier. This work is a first contribution in this direction. We concentrate on the analysis of two-dimensional sequential forward feature selection methods, encompassing a total of eight methods. For the analysis, we define a scenario with two classification classes and a set of representative features (relevant, redundant, and irrelevant), linearly related with the classes, which was carefully designed to bring out differences among the methods and situations where the methods may not perform correctly. A similar, but not completely coincident, scenario was considered by other authors [21, 25], but our analysis proceeds theoretically, to determine the true feature ordering for the methods under analysis. The ordering obtained in this way does not depend on entropy or MI estimation methods, classifiers, or specific datasets, leading to an undoubtful comparison of the feature selection methods, which is the major advantage of our approach. Besides providing an evaluation independent from the classifier, our theoretical framework also unveils several problems intrinsic to the methods, difficult to detect through an evaluation strictly based on data. In particular, we detected inconsistencies in the construction of the objective function used to select the candidate features, due to various types of indeterminations and due to the possibility of the entropy of continuous random variables taking null and negative values.

In Section 2 we review the notions of entropy and MI, as well as their properties, highlighting differences of these notions in the context of discrete and continuous distributions. Section 3 presents the concepts of relevance and redundancy, introducing the idea of relevance-optimal sets. Section 4 surveys the feature selection methods that are evaluated in this work. Then, in Section 5, we propose an evaluation scenario, and for that scenario derive, in Section 6, the theoretical expressions of the entropies and MI that are required to obtain the true feature ordering. In Section 7, we compare the methods under evaluation based on the true feature ordering, and discuss the shortcomings resulting from the possibility of having negative entropies and indeterminations in their objective functions. In Section 8 we present a simulation study on the estimation of the feature ordering which corroborates the theoretical results. Finally, in Section 9 we draw the main conclusions of the work.

2. MUTUAL INFORMATION AND ENTROPY

MI is a measure of association between variables, capturing both linear and non-linear dependencies, that has gained wide acceptance [19]. The MI between two discrete random variables X and Y , denoted $\text{MI}(X, Y)$, is defined by

$$(1) \quad \text{MI}(X, Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{XY}(x, y) \ln \frac{p_{XY}(x, y)}{p_X(x)p_Y(y)}$$

where \mathcal{Z} and $p_Z(\cdot)$ denote the support and probability function of a discrete random variable or random vector Z , with the convention that $0 \ln 0 = 0$. It follows from the definition that

- (a) $\text{MI}(X, Y) = \text{MI}(Y, X)$.
- (b) $\text{MI}(X, Y) \geq 0$, with equality for independent random variables.

The MI between X and Y can also be written in terms of the entropies of X and Y . The entropy of a discrete random variable Z , $H(Z)$, is a measure of the uncertainty of Z and is given by

$$(2) \quad H(Z) = - \sum_{z \in \mathcal{Z}} p_Z(z) \ln p_Z(z).$$

The definition of entropy can be extended to two or more discrete random variables. For the case of two discrete random variables X and Y , the entropy of (X, Y) is defined by:

$$(3) \quad H(X, Y) = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{XY}(x, y) \ln p_{XY}(x, y).$$

It follows that entropy is a nonnegative function, which is null only for degenerate (point mass) random variables or vectors. After performing simple analytical manipulations, one may conclude that

- (c) $\text{MI}(X, X) = H(X)$;
- (d) $\text{MI}(X, Y) = H(X) + H(Y) - H(X, Y)$;
- (e) $\text{MI}(X, Y) = H(X) - H(X|Y)$;

where the conditional entropy of X given Y , $H(X|Y)$, is given by

$$(4) \quad H(X|Y) = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{XY}(x, y) \ln p_{X|Y=y}(x).$$

The MI between continuous random variables X and Y is defined similarly to the case of discrete random variables. In detail, if we let $f_Z(\cdot)$ denote the probability density function of a random variable or random vector Z , then for (X, Y) (absolutely) continuous:

$$(5) \quad \text{MI}(X, Y) = \int_{\mathcal{Y}} \int_{\mathcal{X}} f_{XY}(x, y) \ln \frac{f_{XY}(x, y)}{f_X(x)f_Y(y)} dx dy.$$

Note that properties (a)-(b) stated above also hold for (absolutely) continuous random pairs (X, Y) . Likewise the MI, the entropy of X , $h(X)$, the entropy of (X, Y) , $h(X, Y)$, and the conditional entropy of X given Y , $h(X|Y)$, are given by:

$$(6) \quad h(X) = - \int_{\mathcal{X}} f_X(x) \ln f_X(x) dx,$$

$$(7) \quad h(X, Y) = - \int_{\mathcal{Y}} \int_{\mathcal{X}} f_{XY}(x, y) \ln f_{XY}(x, y) dx dy,$$

$$(8) \quad h(X|Y) = - \int_{\mathcal{Y}} \int_{\mathcal{X}} f_{XY}(x, y) \ln f_{X|Y=y}(x) dx dy.$$

In the remainder of this section, we will use the common terminology of calling differential entropy the entropy function for continuous random variables, $h(\cdot)$. In the paper, we will drop the term “differential” whenever it is clear that we are referring to continuous random variables.

It is important to note that entropy and differential entropy do not share the same properties, even though properties (d)-(e) above hold with entropy substituted by differential entropy. For example, contrarily to the entropy, which is always nonnegative, the differential entropy can take both positive and negative values, as well as zero. This fact is nicely illustrated, for example, by the uniform distribution on the interval $[0, a]$, $a > 0$, $\text{Unif}(0, a)$, for which

$$X \sim \text{Unif}(0, a) \implies h(X) = \ln a.$$

Thus, $h(X)$ is positive (null, negative) if $a > 1$ ($a = 1$, $a < 1$). Note, in particular, that the uniform distribution on the interval $[0, 1]$ has null differential entropy despite this distribution not being close to a degenerate one. Another property that is not shared by entropy and differential entropy is $\text{MI}(X, X)$ being equal to the entropy of X , which we have seen to hold for a discrete random variables [see property (c) above]. A very different result is obtained in the case of an absolutely continuous random variable X , namely that $\text{MI}(X, X) = +\infty$, as stated in [26] and [27].

In practice, it is also important to compute the MI between discrete and continuous random variables. This is the case, for example, in feature selection problems involving a continuous candidate feature and a discrete class variable. For continuous X and discrete Y random variables, the MI between X and Y is given by

$$(9) \quad \text{MI}(X, Y) = \sum_{y \in \mathcal{Y}} \int_{\mathcal{X}} f_{X|Y=y}(x) p_Y(y) \ln \frac{f_{X|Y=y}(x)}{f_X(x)} dx.$$

One may note that properties (a)-(b) and the analogous of properties (d)-(e) still hold in this case.

When dealing with more than two variables, it arises the need to compute the MI among three or more variables. One of the main definitions of MI among three variables is the triple mutual information, TMI [28]. For example, the TMI among continuous random variables X, Y, Z , with joint probability density function $f(x, y, z)$, and marginal distributions $f_{XY}(x, y)$, $f_{XZ}(x, z)$, $f_{YZ}(y, z)$, $f_X(x)$, $f_Y(y)$, and $f_Z(z)$ is given by

$$\text{TMI}(X, Y, Z) = \int_{\mathcal{Z}} \int_{\mathcal{Y}} \int_{\mathcal{X}} f(x, y, z) \ln \frac{f_{XY}(x, y) f_{XZ}(x, z) f_{YZ}(y, z)}{f(x, y, z) f_X(x) f_Y(y) f_Z(z)} dx dy dz.$$

Using this definition, we can prove that for random variables X and Y and a random variable or random vector \mathbf{Z} :

$$(10) \quad \begin{aligned} \text{TMI}(X, Y, \mathbf{Z}) &= \text{MI}(X, Y) - \text{MI}(X, Y|\mathbf{Z}) \\ &= \text{MI}(X, \mathbf{Z}) - \text{MI}(X, \mathbf{Z}|Y) \\ &= \text{MI}(Y, \mathbf{Z}) - \text{MI}(Y, \mathbf{Z}|X), \end{aligned}$$

where

$$\text{MI}(X, Y|\mathbf{Z}) = h(X|\mathbf{Z}) - h(X|Y, \mathbf{Z}),$$

and analogously for the other cases. The definition of $\text{TMI}(X, Y, \mathbf{Z})$ has the disadvantage of assuming not only positive or null values but also negative ones [28], which demands a new interpretations of MI.

3. RELEVANCE AND REDUNDANCY

Feature selection methods share the general goal of identifying an appropriate subset of the original features with the property of being *maximally informative* about the class [8, 25]. Following the principle of parsimony, among the maximally informative sets the ones that have minimum size are to be preferred; we call these minimum size sets *relevance-optimal* sets.

In order to introduce the notion of maximally informative and relevance-optimal sets, it is convenient to introduce some notation. We let $\mathbf{V} = (V_i)_{i \in T}$ denote the set of all input features and C the class (random

variable). Moreover, for a subset L of T we let $\mathbf{V}_L = (V_l)_{l \in L}$, and similarly for an observation $\mathbf{v} = (v_i)_{i \in T}$ of $\mathbf{V} = (V_i)_{i \in T}$ we let $\mathbf{v}_L = (v_l)_{l \in L}$. In addition, we let $\bar{L} = T \setminus L$ denote the complement of L , $T_{-i} = T \setminus \{i\}$, and $\stackrel{d}{=}$ denote equality in distribution.

Definition 1. The feature set \mathbf{V}_L is maximally informative (for class C) if for all \mathbf{v} in the support of \mathbf{V} ,

$$[C | (\mathbf{V}_L, \mathbf{V}_{\bar{L}}) = (\mathbf{v}_L, \mathbf{v}_{\bar{L}})] \stackrel{d}{=} [C | \mathbf{V}_L = \mathbf{v}_L].$$

Moreover, a maximally informative (feature) set is a relevance-optimal set if it has minimum size among all maximally informative sets.

Thus, a feature set \mathbf{V}_L is maximally informative for class C if knowledge on features not belonging to \mathbf{V}_L does not impact the conditional distribution of C , provided the values of the features belonging to \mathbf{V}_L are known. With the previous definition, we are in condition to introduce the concept of *irrelevant* feature, as well as two concepts of feature relevance: *strongly relevant* feature and *weakly relevant* feature.

Definition 2. A feature V_i is strongly relevant if $\mathbf{V}_{T_{-i}}$ is not maximally informative and is irrelevant if for all $L \subseteq T_{-i}$ and all (v_i, \mathbf{v}_L) in the support of (V_i, \mathbf{V}_L) :

$$[C | (V_i, \mathbf{V}_L) = (v_i, \mathbf{v}_L)] \stackrel{d}{=} [C | \mathbf{V}_L = \mathbf{v}_L].$$

A feature that is neither strongly relevant nor irrelevant is called *weakly relevant*.

The previous definition leads to a partition of the set of features into strongly relevant (SR), weakly relevant (WR), and irrelevant features, with the definitions of SR, WR, and irrelevant features coinciding with the ones presented in [12] and [15]. For a characterization of SR feature, WR feature, and irrelevant feature based on TMI see [7].

Note that a SR feature belongs to all relevance-optimal sets. Conversely, an irrelevant feature belongs to no relevance-optimal set. Furthermore, a relevance-optimal set may either contain or do not contain a specific WR feature. Thus, in general the identification of relevant (SR and WR) features is not enough to get a relevance-optimal subset, since duplications or other kinds of functional dependencies may occur among WR features. The next example, which is inspired in Example 1 of [15], illustrates this fact.

Example 1. Let $T = \{1, 2, \dots, 5\}$, with features V_1 , V_2 , and V_4 being independent, $V_3 = 3V_2 + 1$, and $V_5 = (V_4)^2$. Moreover, assume that $C = g(V_1, V_2)$ is a binary random variable such that $g(v_1, v_2)$ is not constant in either of the variables v_1 and v_2 .

The sets containing feature V_1 and one of the features V_2 and V_3 are maximally informative since both (V_1, V_2) and (V_1, V_3) determine C , whereas none of the features V_1 , V_2 , and V_3 in isolation determines C . As a by-product, we conclude that: V_1 is the unique SR feature; V_2 and V_3 are WR features; and V_4 and V_5 are irrelevant features. Moreover, there are two relevance-optimal sets: (V_1, V_2) and (V_1, V_3) .

Note that the relevance-optimal sets have size two, thus implying that a minimum of two features are needed to convey all information on the class that is contained in (V_1, V_2, \dots, V_5) .

The scientific community quickly realized that, for the large and complex feature sets commonly found in practice, it may be impractical to derive all relevance-optimal sets. This has paved the way to the development of the systematic approach to derive (just) a single relevance-optimal set using Markov blanket filtering [13].

Markov blanket filtering is a backward elimination process that starting from the set of relevant (SR and WR) features, say \mathbf{V}_R , eliminates one by one WR features until a relevance-optimal set is obtained. Each step of the backward elimination process consists in selecting a feature V_j from the current maximally-relevant subset \mathbf{V}_N of \mathbf{V}_R for which there exists a Markov blanket \mathbf{V}_M , $M \subseteq N \setminus \{j\}$, meaning that for any (v_j, \mathbf{v}_M) in the support of (V_j, \mathbf{V}_M) ,

$$[(C, \mathbf{V}_K) | (V_j, \mathbf{V}_M) = (v_j, \mathbf{v}_M)] \stackrel{d}{=} [(C, \mathbf{V}_K) | \mathbf{V}_M = \mathbf{v}_M].$$

with $K = N \setminus (M \cup \{j\})$. Following this way, a relevance-optimal set is obtained when none of the features $V_j, j \in N$, possesses a Markov blanket.

Yu and Liu in [15] rightly pointed out that we cannot find a Markov blanket for strongly relevant features, thus implying that a relevance-optimal set contains necessarily all SR features. However, a relevance-optimal set contains only a part of the WR features, as illustrated in the example above. As a result, each relevance-optimal set leads naturally to the following classification of WR features in two types: WR features that belong to the relevance-optimal set, and WR features that do not belong to the relevance-optimal set (see [12, 15], and references therein). Following [15], we call the former weakly relevant and non-redundant (WR-NR) features and the latter weakly relevant and redundant (WR-R) features. As a result, one gets a partition of the set of features in four subsets: SR features, WR-NR features, WR-R features, and irrelevant features.

One should stress that the partition of the features in four sets (SR, WR-NR, WR-R, and irrelevant) thus obtained is a function of the relevance-optimal set used to divide WR features into WR-NR and WR-R features. As an illustration, note that in Example 1 the feature V_2 is WR-NR for the relevance-optimal set (V_1, V_2) and WR-R for the relevance-optimal set (V_1, V_3) , with the converse holding for feature V_3 .

To end the section, we remark that not all relevance-optimal sets should be considered equally good from a practical point of view. In fact, the degrees of redundancy (or association) between the features of different relevance-optimal sets are not necessarily equal, in which case relevance-optimal sets whose features exhibit the lowest level of redundancy should be preferred. This may be interpreted as a reason for selecting a relevance-optimal minimum-redundancy feature set.

4. FEATURE SELECTION METHODS

In feature selections problems, the practitioner aims at finding features that contain as much information as possible about the class variable while, following the principle of parsimony and seeking to improve interpretability, avoid selecting features that contain redundant information with respect to the class variable. We will concentrate in the framework of forward sequential methods that progressively select a new feature, to add to the set of selected features, using criteria based on MI measures.

With S (F) denoting the set of already-selected (unselected) features at a given step, $F \cap S = \emptyset$, the concrete objective turns out to be the selection of an additional feature $V_i \in F$ such that the MI between the class variable (C) along with the already-selected features (S) and the candidate feature (V_i), $\text{MI}(C, \{S, V_i\})$, is maximized. However,

$$\begin{aligned} \text{MI}(C, \{S, V_i\}) &= \text{MI}(C, S) + \text{MI}(C, V_i) - \text{TMI}(C, V_i, S) \\ &= \text{MI}(C, S) + \text{MI}(C, V_i|S), \end{aligned}$$

and $\text{MI}(C, S)$ does not depend on i , this amounts to pick a feature $V_i \in F$ such that the conditional MI between V_i and the class variable given the set of already-selected features, $\text{MI}(C, V_i|S)$, is maximized. Here $\text{TMI}(C, V_i, S)$ represents the redundancy between the candidate feature V_i , the already-selected features S , and the class variable C . As a result, the decomposition of $\text{MI}(C, V_i|S)$ expresses a trade off between the *relevance* of the candidate feature to explain the class variable, encompassed in $\text{MI}(C, V_i)$, and its *redundancy* for the same effect in face of the previously selected features, $\text{TMI}(C, V_i, S)$.

One should note that in practice the estimation of $\text{TMI}(C, V_i, S)$ is problematic even in the case where S contains a small number of features [21, 22, 25]. Accordingly, several feature selection methods that use simplifications to approximate $\text{TMI}(C, V_i, S)$ have been introduced in the literature. In this paper, we only study methods that consider a two-dimensional approximation of the TMI (for three-dimensional alternatives see [7–11] and references therein). These feature selection methods include: MIFS [20], MIFS-U [21], mRMR [22], mMIFS-U [24], MICC [25], QMIFS [25], and NMIFS [23]. In brief, these methods select

at each step a feature according to the following type of criteria

$$(11) \quad \arg \max_{V_i \in F} \{ \text{MI}(C, V_i) - \overline{\text{TMI}}(C, V_i, S) \}$$

where $\overline{\text{TMI}}(C, V_i, S)$ is a method dependent approximation to $\text{TMI}(C, V_i, S)$. We next present the specific forms of $\overline{\text{TMI}}(C, V_i, S)$ for the mentioned methods. The objective functions for the methods are summarized in Table 1.

TABLE 1. Objective functions of the eight different feature selection methods: MIFS, MIFS-U, mRMR, mMIFS-U, QMIFS, MICC, NMIFS and maxMIFS.

Method	Objective function
MIFS [20]	$\arg \max_{V_i \in F} \left\{ \text{MI}(C, V_i) - \beta \sum_{V_s \in S} \text{MI}(V_i, V_s) \right\}$
MIFS-U [21]	$\arg \max_{V_i \in F} \left\{ \text{MI}(C, V_i) - \beta \sum_{V_s \in S} \frac{\text{MI}(C, V_s)}{h(V_s)} \text{MI}(V_i, V_s) \right\}$
mRMR [22]	$\arg \max_{V_i \in F} \left\{ \text{MI}(C, V_i) - \frac{1}{ S } \sum_{V_s \in S} \text{MI}(V_i, V_s) \right\}$
mMIFS-U [24]	$\arg \max_{V_i \in F} \left\{ \text{MI}(C, V_i) - \max_{V_s \in S} \frac{\text{MI}(C, V_s)}{h(V_s)} \text{MI}(V_i, V_s) \right\}$
MICC [25]	$\arg \max_{V_i \in F} \left\{ \frac{\text{MI}(C, V_i)}{\frac{1}{ S } \sum_{V_s \in S} \frac{\text{MI}(V_i, V_s)}{\min\{h(V_i), h(V_s)\}}} - \text{MI}(C, V_i) \right\}$
QMIFS[25]	$\arg \max_{V_i \in F} \left\{ \text{MI}(C, V_i) - \sum_{V_k \in S} \left(\frac{\text{MI}(V_i, V_k)}{h(V_k)} - \frac{1}{2} \sum_{\substack{V_j \in S \\ j \neq k}} \frac{\text{MI}(V_i, V_j)}{h(V_j)} \frac{\text{MI}(V_j, V_k)}{h(V_k)} \text{MI}(C, V_k) \right) \right\}$
NMIFS [23]	$\arg \max_{V_i \in F} \left\{ \text{MI}(C, V_i) - \frac{1}{ S } \sum_{V_s \in S} \frac{\text{MI}(V_i, V_s)}{\min\{h(V_i), h(V_s)\}} \right\}$
maxMIFS [29]	$\arg \max_{V_i \in F} \left\{ \text{MI}(C, V_i) - \max_{V_s \in S} \{ \text{MI}(V_i, V_s) \} \right\}$

The first proposal, developed in [20] and called MIFS method, uses the approximation

$$(12) \quad \overline{\text{TMI}}(C, V_i, S) = \beta \sum_{V_s \in S} \text{MI}(V_i, V_s)$$

where $\beta \in [0, 1]$ is a weight factor that should be chosen by the user. One may arrive at this approximation by introducing a weight factor β after initially assuming that:

(a) $\text{TMI}(C, V_i, S) = \text{MI}(V_i, S)$.

(b) The already-selected features are independent.

Battiti's [20] first assumption states that, given a certain class $C = c$, the candidate feature, V_i , and the already-selected features, S , are independent, an hypothesis of conditional independence. The assumptions (a)-(b) lead to

$$\text{TMI}(C, V_i, S) = \text{MI}(V_i, S) = \sum_{V_s \in S} \text{MI}(V_i, V_s).$$

The introduction of the weight factor β may thus be regarded as a correction factor for deviations from the two mentioned assumptions. This parameter was viewed by the author of MIFS as regulating the relative importance of the redundancy component. [20] claimed that a value for β [in (12)] between 0.5 and 1 is appropriate for many classification tasks. However, several authors have argued that the best choice for β being problem dependent constitutes an important drawback of MIFS.

The mRMR method, proposed in [22], avoids the need to choose a value for the parameter β . Even though it has been derived by its authors as a criteria combining maximum relevance with minimum redundancy, the mRMR method corresponds to a variation of the MIFS method through the introduction of an adaptive β that evolves as the number of already-selected features changes, being effectively the reciprocal of the the number of already-selected features. More precisely, mRMR uses the approximation

$$(13) \quad \overline{\text{TMI}}(C, V_i, S) = \frac{1}{|S|} \sum_{V_s \in S} \text{MI}(V_i, V_s).$$

Note that the redundancy component associated with the selection of candidate feature V_i is here measured by the mean of the MI between V_i and each of the already-selected features, $V_s \in S$.

With the aim of addressing the fact that the entropy of random variables may vary greatly, [23] claims that the MI values between the candidate feature and the already-selected features should be normalized. Accordingly, its authors proposed to substitute $\text{MI}(V_i, V_s)$ by $\text{NI}(V_i, V_s)$, the normalized mutual information between the features V_i and V_s , given by

$$(14) \quad \text{NI}(V_i, V_s) = \frac{\text{MI}(V_i, V_s)}{\min\{h(V_i), h(V_s)\}}.$$

In sequence, they proposed in the same paper the NMIFS method, which uses the approximation

$$(15) \quad \overline{\text{TMI}}(C, V_i, S) = \frac{1}{|S|} \sum_{V_s \in S} \text{NI}(V_i, V_s).$$

The use of $\text{NI}(V_i, V_s)$ in NMIFS – instead of $\text{MI}(V_i, V_s)$, like in mRMR – as a measure of redundancy between the candidate feature V_i and the already-selected feature V_s was justified with the supposed fact that

$$0 \leq \text{MI}(V_i, V_s) \leq \min\{h(V_i), h(V_s)\}$$

leading to $0 \leq \text{NI}(V_i, V_s) \leq 1$. However, the second inequality in the above equation only holds with certainty when V_i and V_s are discrete random variables. In fact, as the entropies of continuous random variables may take negative values, $\text{NI}(V_i, V_s)$ can take negative values, leading to the redundancy of V_i with respect to V_s being weighted positively, contrarily to what was intended. This problem extends to all other methods that incorporate entropies of features in denominators of fractions.

Kwak and Choi [21] introduced the MIFS-U method, whose basis is similar to that of MIFS, but where the authors tried to overcome the assumption of independence between the class and the redundancy component - assumption (a), while maintaining the independence assumption for the already-selected features - assumption (b). Specifically, MIFS-U uses the approximation

$$(16) \quad \overline{\text{TMI}}(C, V_i, S) = \beta \sum_{V_s \in S} \frac{\text{MI}(C, V_s)}{h(V_s)} \text{MI}(V_i, V_s)$$

by assuming that the class variable does not change the ratio of the MI of the candidate feature with a single already-selected feature to the entropy of that already-selected feature, i.e.,

$$(17) \quad \frac{\text{MI}(V_i, V_s)}{h(V_s)} = \frac{\text{MI}(V_i, V_s|C)}{h(V_s|C)}$$

for each $V_s \in S$. A direct consequence of this assumption is that

$$\begin{aligned} \text{TMI}(C, V_i, V_s) &= \text{MI}(V_i, V_s) - \text{MI}(V_i, V_s|C) \\ &= \left[1 - \frac{h(V_s|C)}{h(V_s)} \right] \text{MI}(V_i, V_s) \\ &= \frac{\text{MI}(C, V_s)}{h(V_s)} \text{MI}(V_i, V_s), \end{aligned}$$

which leads to the term that appears in (16), the MIFS-U approximation for $\text{MI}(C, V_i, S)$.

The assumption (17) is somehow counterintuitive as one expects that if features are associated with the class variable, then knowledge of the class variable would lead to different conditional information on the features. Moreover the appearance of the entropies of the already-selected features in the denominators of fractions in (16) constitutes a drawback of MIFS-U in the presence of already-selected features with entropy close to zero, and especially in the presence of already-selected continuous-type features with negative entropy. As a result, the approximation (16) for $\text{TMI}(C, V_i, S)$ may turn out to be negative, leading to the redundancy of the candidate feature with already-selected features being weighted positively, contrarily to what was desired.

Novovicová and co-authors [24] proposed the mMIFS-U method, which uses the approximation

$$(18) \quad \overline{\text{TMI}}(C, V_i, S) = \max_{V_s \in S} \left\{ \frac{\text{MI}(C, V_s)}{h(V_s)} \text{MI}(V_i, V_s) \right\}.$$

Like the MIFS-U method, mMIFS-U assumes the condition (17), and shares with MIFS-U the drawbacks resulting from having entropies of already-selected variables appearing in the denominators of fractions. Conversely, contrarily to MIFS-U, mMIFS-U avoids the problem of selecting an appropriate value for β by replacing a sum over the already-selected features in (16) by a maximum over the same set of features in (18).

Later, [25] introduced the QMIFS method, which uses the following approximation for $\text{MI}(C, V_i, S)$ with the aim of incorporating possible interactions between two (but not more than two) already-selected features:

$$(19) \quad \overline{\text{MI}}(C, V_i, S) = \sum_{V_k \in S} \left[\phi_{ik} - \frac{1}{2} \sum_{\substack{V_j \in S \\ j \neq k}} \phi_{ij} \phi_{jk} \right] \text{MI}(C, V_k)$$

where $\phi_{lm} = \text{MI}(V_l, V_m)/h(V_m)$ for $l \neq m$. The presentation of the authors for the derivation of this approximation is not easy to follow and seems to be based on several assumptions that may be hardly satisfied in practice. In particular, aside from the condition (17), the authors assume the following property on the information of the candidate feature V_i and pairs of already-selected features (V_j, V_k) , with $V_j \neq V_k$:

$$(20) \quad \frac{\text{MI}(V_i, V_j)}{h(V_j)} = \frac{\text{MI}(V_i, V_j, V_k)}{\text{MI}(V_j, V_k)}.$$

As a result of (19) and what has been stated, one concludes that QMIFS shares the main drawbacks of MIFS-U not related with the parameter β of the latter.

In this paper we consider an additional feature selection method, which we call maxMIFS. This method is similar to mRMR [22], but uses the maximum MI between the candidate feature and individual already-selected features instead of their mean. That is, maxMIFS is a method of the generic type criteria (11) with

the approximation

$$(21) \quad \overline{\text{TMI}}(C, V_i, S) = \max_{V_s \in S} \text{MI}(V_i, V_s).$$

Note that the use of the maximum of the MI between the candidate feature and each of the already-selected features avoids overweighting the redundancy component of the objective function.

To end the section, one should mention that [25] proposed a second feature selection method called MICC. Like NMIFS, this method is based on the use of the normalized mutual information between the candidate features V_i and already-selected V_s , $\text{NI}(V_i, V_s)$, instead of the MI between the same variables. In detail, MICC uses as criteria for selecting a new feature the candidate feature V_i that maximizes the following expression:

$$(22) \quad \overline{\text{TMI}}(C, V_i | S) = \text{MI}(C, V_i) \left[\frac{|S|}{\sum_{V_s \in S} \text{NI}(V_i, V_s)} - 1 \right].$$

Similarly to the reasoning followed in the proposal of the NMIFS method [23], it is claimed in [25] that the multiplicative factor affecting $\text{MI}(C, V_i)$ in the previous equation takes values on $[0, \infty]$. However, this conclusion may be false for continuous features.

5. EVALUATION SCENARIO

Evaluating feature selection methods can be done in two ways. The first one is to embed the classifier in the evaluation process [20, 22, 24]. In this case, the methods are compared based on the accuracy of the classification process estimated using labeled data (data for which the true class is known). The results obtained with this method are difficult to generalize, since they depend on the specific classifier and on the performance metrics used in the comparison. The second evaluation method is based on scenarios defined by an initial set of *interesting* features and a relation between these features and the output class [21, 25]. In this case, the true ordering of features must be known, and the methods are compared based on how well they can approach it. A reference that may be used in this type of evaluation is the one obtained with the Markov blanket filtering methodology described in Section 3. In this work we will concentrate on the last type of evaluation.

There are three requirements that a good evaluation scenario must observe. First, it must be challenging, i.e., it must lead to situations where the decision metrics used in selecting candidate features are close enough to favor wrong decisions. Second, it must include a representative set of features, containing relevant, redundant, and irrelevant ones. Finally, it must be amenable to theoretical evaluation, i.e., one should be able to obtain the *true* ordering of features for the methods under analysis. If this last requirement is not fulfilled, the evaluation can only be based on a conjecture of what the true ordering is, which may lead to erroneous evaluation. To the best of our knowledge, our work is the first one to utilize a theoretical framework in the evaluation of feature selection methods.

Kwak and Choi [21] and Huang et al. [25] proposed an evaluation scenario with two classes defined by

$$(23) \quad C = \begin{cases} 0, & X + 0.2Y < 0 \\ 1, & X + 0.2Y \geq 0 \end{cases}$$

where X and Y are independent random variables uniformly distributed in $[-0.5, 0.5]$. [21] considered as interesting features X , Y^2 , and $X - Y$; [25] added seven other features, including Z and functions of it, where Z is independent and identically distributed to X and Y (see Table 2 for the complete list). Both evaluation scenarios are amenable to theoretical evaluation, as it will become clear in next section, but the authors did not pursue this goal.

Using the framework of Section 3, the features in [25] can be classified in the following way: there are no strongly relevant features; Z and Z^2 are irrelevant; Y^2 and ZXY are always WR-R features. Moreover,

TABLE 2. Input features proposed in [25].

Features	Description	Features	Description
V_1	X	V_6	Z^2
V_2	$3X + 1$	V_7	ZX
V_3	Y^2	V_8	ZY
V_4	$X - Y$	V_9	ZXY
V_5	Z	V_{10}	ZY

there are two relevance-optimal sets: $(X, X - Y)$ and $(3X + 1, X - Y)$. This selection of features deserves the following comments:

- It is reasonable that no strongly relevant feature has been included, since these have a high probability of being selected as relevant. SR features do not put the feature selection method under stress.
- The number of interesting relevance-optimal sets is too small. In fact, as discussed in Section 4, the methods under analysis perform selection by evaluating the relevance of features (to the class) and the redundancy between the candidate and already-selected features. Thus, it is important to include in the initial set features that lead to relevance-optimal sets with different levels of redundancy among features.
- Strangely, Y was not included in the set of features, given that it is one of the features used in the class definition. Including Y would have added two relevance-optimal sets, (X, Y) and $(3X + 1, Y)$, where features are independent among themselves. Moreover, to evaluate how well the feature selection methods match the true feature ordering, it is important to confront the possibility of selecting (X, Y) or $(X, X - Y)$, or equivalently $(3X + 1, Y)$ and $(3X + 1, X - Y)$. These two outcomes are easily confused. Indeed, as we will show later, $X - Y$ has a MI with the class which is larger than that of Y . Thus, depending on the relative strength of the redundancy component, either (X, Y) or $(X, X - Y)$ may be selected first.

Based on the above comments, we generalized the evaluation scenario of [25], in the following way. First, we included Y in the set of features. Second, we removed features ZX , ZY , and ZXY , because theoretical analysis is involved and these are necessarily WR-R features. Finally, we added two irrelevant features W and $W + Z$, to assess whether the feature selection methods lead to particular patterns of feature ordering (e.g. irrelevant or redundant features following the relevance-optimal set). Our scenario is then based on the 10 features shown in Table 3. We also expanded the class definition, to contemplate different relative strengths between X and Y . Specifically, the two classes are defined by

$$C_k = \begin{cases} 0, & X + kY < 0 \\ 1, & X + kY \geq 0 \end{cases},$$

where $k \in [0, 1]$. In this way, our scenario has four irrelevant features, Z , Z^2 , $W + d$, and $W + Z$, no strongly relevant feature, two features that are WR-R, X^2 and Y^2 , and five relevance-optimal sets, (X, Y) , $(X, X - Y)$, $(Y, X - Y)$, $(3X + 1, Y)$, and $(3X + 1, X - Y)$.

TABLE 3. Input features of evaluation scenario.

Features	Description	Features	Description
V_1	X	V_6	Z^2
V_2	$aX + b$	V_7	Y
V_3	Y^2	V_8	X^2
V_4	$X - Y$	V_9	$W + d$
V_5	Z	V_{10}	$Z + W$

6. THEORETICAL ENTROPY AND MUTUAL INFORMATION

In this section we summarize the theoretical results needed to compare the feature selection methods. We consider two different scenarios, where the random variables X , Y , Z , and W are considered independent and identically distributed. In Scenario I, the random variables follow a uniform distribution on $[-\delta, \delta]$, and in Scenario II a standard normal distribution, $\mathcal{N}(0, 1)$.

Given the extensive derivations needed to prove the results we have established, we only highlight in this section the less intuitive or most relevant aspects. The complete derivations can be found in [29].

The theoretical evaluation of the feature selection methods, whose objective functions are summarized in Table 1, need the following expressions:

- (i) Entropy of the class, C_k , and of all features, V_i , $i = 1, \dots, 10$ (see Tables 4 and 5 for Scenarios I and II, respectively).
- (ii) MI between the class and each feature, $\text{MI}(C_k, V_i)$, $i = 1, \dots, 10$ (see also Tables 4 and 5 for Scenarios I and II, respectively).
- (iii) MI between each pair of features, $\text{MI}(V_i, V_j)$, $i, j \in \{1, \dots, 10\}$ (see Table 6 for Scenario I and Table 7 for Scenario II).

From Table 4, we realise that if $X \sim \text{Unif}(-\delta, \delta)$, the entropy of X is $h(X) = \ln(2\delta)$. This is a known result [19]. Note, however, that if $\delta = 0.5$, $h(X) = 0$ and if $0 < \delta < 0.5$ then $h(X) < 0$, which stresses the fact that the entropies of continuous and discrete features do not have the same properties and require different interpretations. In next section we are going to show the impact of this fact in the performance of feature selection methods. Given its complexity, the general expression for $\text{MI}(C_k, X)$ is only defined for $\delta \geq 0.5$; its general form is provided in [29].

With the exception of $Y^2 \stackrel{d}{=} X^2$ and $X - Y$, the derivation of the density functions of the features in Scenario I is quite simple. For the first case, one has

$$(24) \quad f_{Y^2}(u) = \begin{cases} \frac{1}{2\delta\sqrt{u}}, & 0 \leq u \leq \delta^2 \\ 0, & \text{elsewhere} \end{cases}.$$

As this is not a commonly known distribution, its entropy is calculated, leading to the expression provided in Table 4. It can be shown that if X and Y are two independent features with $\text{Unif}(-\delta, \delta)$ then $X - Y$ has a Triangular distribution with lower limit -2δ , upper limit 2δ and mode 0, *i.e.*, $X - Y \sim \text{Tri}(-2\delta, 2\delta, 0)$. The entropy of this triangular distribution is known and is provided in Table 4. For simplicity, in Table 4 we present $\text{MI}(C_k, X - Y)$ only for the case when $\delta = 0.5$, value suggested in [21] and [25]; the general expression of this entropy is available in [29].

A non-intuitive result, true for scenarios I and II, is that $\text{MI}(C_k, X^2) = 0$ (proof provided in Appendix A), meaning that even though X is important in the definition of C_k , X^2 has no association with the class. Another relevant fact is that, for Scenario I, $\text{MI}(C_k, Y)$ does not depend on δ .

In Scenario II, the features X , Y , Z and W have standard normal distribution, whose known good properties guarantee that all features under study have known distributions. Nevertheless, we raise the attention to X^2 (similarly Y^2 and Z^2) which have chi-squared distribution with one degree of freedom, $\chi^2_{(1)}$. It is known that the entropy of a random variable with chi-squared distribution with k degrees of freedom is $k/2 + \ln(2\Gamma(k/2)) + (1 - k/2)\psi(k/2)$ where $\Gamma(\cdot)$ is the Gamma function and $\psi(\cdot)$ is the Digamma function [30].

The calculation of MI between each feature and the class (apart from cases of independence) requires the use of the family of univariate skew normal distributions. This family generalizes the normal distribution, allowing skewness different from zero [31]. A feature V with skew normal distribution with location $\mu \in \mathbb{R}$, scale $\sigma > 0$, and shape $\alpha \in \mathbb{R}$ is represented by $\text{SN}(\mu, \sigma, \alpha)$. Note that, $(V - \mu)/\sigma \sim \text{SN}(0, 1, \alpha)$, and if $\alpha = 0$ then $V \sim \mathcal{N}(\mu, \sigma)$.

We recall that the MI is symmetric and non-negative (being 0 for independent features), is invariant under scale or location transformations and under one-to-one transformations (that is, $\text{MI}(X, Y) = \text{MI}(U, W)$ where $u = g(x)$, $w = g(y)$, and g is invertible, see [32] for details). These properties justify the following results:

- (i) All zeros in Tables 6 and 7;
- (ii) $\text{MI}(X, X - Y) = \text{MI}(aX + b, X - Y)$
 $= \text{MI}(Y, X - Y) = \text{MI}(Z, W + Z)$
 $= \text{MI}(W + d, W + Z)$, and
- (iii) $\text{MI}(X^2, X - Y) = \text{MI}(Y^2, X - Y) = \text{MI}(Z^2, W + Z)$.

Once again, it can be proved (vide [29] for details) that:

$$\text{MI}(X, X - Y) = \begin{cases} \frac{1}{2}, & \text{Scenario I} \\ \frac{\ln 2}{2}, & \text{Scenario II} \end{cases}.$$

In a similar manner, it can be established that

$$\text{MI}(Y^2, X - Y) = \begin{cases} \frac{1 - \ln 2}{2}, & \text{Scenario I} \\ -1 + \frac{\ln 2}{2} + E(\ln(\cosh[(X - Y)|Y|])), & \text{Scenario II.} \end{cases}$$

For Scenario II, the expression for $\text{MI}(Y^2, X - Y)$ is evaluated numerically, using the software Mathematica [33], leading to the approximation 0.1078, value included in Table 7.

Even though the MI between two identical discrete features is equal to the entropy of the feature, this property does not hold for absolute continuous features. In fact, [26] proved that if V_i and V_j are two absolute continuous features, where V_j is a measurable function of V_i , then $\text{MI}(V_i, V_j) = +\infty$. This leads to:

- (i) $\text{MI}(V_i, V_i) = +\infty$, for $i = 1, \dots, 10$ and
- (ii)

$$\begin{aligned} \text{MI}(X, aX + b) &= \text{MI}(aX + b, X^2) = \text{MI}(X, X^2) = \text{MI}(Y, Y^2) \\ &= \text{MI}(Z, Z^2) = +\infty. \end{aligned}$$

All these results are summarized in Tables 6 and 7.

TABLE 4. General expression for the entropy of each feature, $h(V_i)$, $i = 1, \dots, 10$, and for the MI between each feature and the class, $\text{MI}(C_k, V_i)$, $i = 1, \dots, 10$, for Scenario I.

Feature	Description	$h(\cdot)$	$\text{MI}(C_k, \cdot)$
V_1	$X \sim \text{Unif}(-\delta, \delta)$	$\ln(2\delta)$	$-\frac{k}{2} + \ln(2)$
V_2	$aX + b \sim \text{Unif}(-\delta a + b, \delta a + b)$	$\ln(2a\delta)$	$\text{MI}(C_k, V_1)$
V_3	Y^2	$\ln(2\delta^2) - 1$	0
V_4	$X - Y \sim \text{Tri}(-2\delta, 2\delta, 0)$	$\frac{1}{2} + \ln(2\delta)$	$\frac{-(k-1)^2 \ln(1-k)}{4k}$
V_5	$Z \stackrel{d}{=} X$	$h(V_1)$	0
V_6	$Z^2 \stackrel{d}{=} Y^2$	$h(V_3)$	0
V_7	$Y \stackrel{d}{=} X$	$h(V_1)$	$\frac{(k^2 + 1) \ln\left(\frac{1+k}{1-k}\right) + 2k(\ln(1-k^2) - 1)}{4k}$
V_8	$X^2 \stackrel{d}{=} Y^2$	$h(V_3)$	0
V_9	$W + 2 \sim \text{Unif}(-\delta + d, \delta + d)$	$h(V_1)$	0
V_{10}	$Z + W \stackrel{d}{=} X - Y$	$h(V_4)$	0

TABLE 5. General expression for the entropy of each feature, $h(V_i), i = 1, \dots, 10$, and for the MI between each feature and the class, $MI(C_k, V_i), i = 1, \dots, 10$, for Scenario II. In $h(V_3)$, γ represents the Euler's constant.

Feature	Description	$h(\cdot)$	$MI(C_k, \cdot)$
V_1	$X \sim \mathcal{N}(0, 1)$	$\frac{1}{2} \ln(2\pi e)$	$E \left[\ln \left(2\Phi \left(-\frac{A_0}{k} \right) \Phi \left(\frac{A_1}{k} \right) \right) \right]$ where $A_0 \sim \text{SN}(0, 1, -1/k)$ and $A_1 \sim \text{SN}(0, 1, 1/k)$
V_2	$aX+b \sim \mathcal{N}(b, a^2)$	$\frac{1}{2} \ln(2\pi e a^2)$	$MI(C_k, V_1)$
V_3	$Y^2 \sim \chi_{(1)}^2$	$\frac{1}{2}(1 + \ln \pi - \gamma)$	0
V_4	$X - Y \sim \mathcal{N}(0, 2)$	$\frac{1}{2} \ln(4\pi e)$	$E \left[\ln \left(2\Phi \left(-\frac{1-k}{k+1} B_0 \right) \Phi \left(\frac{1-k}{k+1} B_1 \right) \right) \right]$ where $B_0 \sim \text{SN} \left(0, 1, -\frac{1-k}{k+1} \right)$ and $B_1 \sim \text{SN} \left(0, 1, \frac{1-k}{k+1} \right)$
V_5	$Y \stackrel{d}{=} X$	$h(V_1)$	0
V_6	$Z^2 \stackrel{d}{=} Y^2$	$h(V_3)$	0
V_7	$Y \stackrel{d}{=} X$	$h(V_1)$	$E[\ln(2\Phi(-kD_0)\Phi(kD_1))]$ where $D_0 \sim \text{SN}(0, 1, -k)$ and $D_1 \sim \text{SN}(0, 1, k)$
V_8	$X^2 \stackrel{d}{=} Y^2$	$h(V_3)$	0
V_9	$W+d \sim \mathcal{N}(d, 1)$	$h(V_1)$	0
V_{10}	$W+Z \sim \mathcal{N}(0, 2)$	$h(V_4)$	0

TABLE 6. MI between the input features, $MI(V_i, V_j), i = 1, \dots, 10, j = 1, \dots, i$, for Scenario I.

$MI(\cdot, \cdot)$	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}
V_1	$+\infty$									
V_2	$+\infty$	$+\infty$								
V_3	0	0	$+\infty$							
V_4	0.5	0.5	$\frac{1-\ln 2}{2}$	$+\infty$						
V_5	0	0	0	0	$+\infty$					
V_6	0	0	0	0	$+\infty$	$+\infty$				
V_7	0	0	$+\infty$	0.5	0	0	$+\infty$			
V_8	$+\infty$	$+\infty$	0	$\frac{1-\ln 2}{2}$	0	0	0	$+\infty$		
V_9	0	0	0	0	0	0	0	0	$+\infty$	
V_{10}	0	0	0	0	0.5	$\frac{1-\ln 2}{2}$	0	0	0.5	$+\infty$

7. THEORETICAL FEATURE ORDERING

In this section, we will present the true feature ordering obtained with the evaluation scenario described in Section 5, and using the results of Section 6. As in [25], we consider for feature $aX + b$ that $a = 3$ and $b = 1$ and that the uniform distribution has parameter $\delta = 0.5$; we also consider that $d = 2$. These concretizations lead to the entropy and MI (with the class) values shown in Table 8 (for scenarios I and II).

The feature ordering results are shown in Table 9 (for Scenario I) and in Table 10 (for Scenario II).

We have seen in previous sections that features can have null and negative entropy, and null MI with the class. It was also seen that the MI between features can be null or $+\infty$. These limiting values have a deep impact in the performance of feature selection methods, a characteristic that seems not have been accounted for in previous works.

TABLE 7. MI between the input features, $MI(V_i, V_j)$, $i = 1, \dots, 10$, $j = 1, \dots, i$, for Scenario II.

$MI(\cdot, \cdot)$	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}
V_1	$+\infty$									
V_2	$+\infty$	$+\infty$								
V_3	0	0	$+\infty$							
V_4	$\frac{\ln 2}{2}$	$\frac{\ln 2}{2}$	0.1078	$+\infty$						
V_5	0	0	0	0	$+\infty$					
V_6	0	0	0	0	$+\infty$	$+\infty$				
V_7	0	0	$+\infty$	$\frac{\ln 2}{2}$	0	0	$+\infty$			
V_8	$+\infty$	$+\infty$	0	0.1078	0	0	0	$+\infty$		
V_9	0	0	0	0	0	0	0	0	$+\infty$	
V_{10}	0	0	0	0	$\frac{\ln 2}{2}$	0.1078	0	0	$\frac{\ln 2}{2}$	$+\infty$

TABLE 8. Entropy of each input variable, $h(V_i)$, $i = 1, \dots, 10$, and MI between each input variable and the class variable, $MI(C_k, V_i)$, $k = 0.2, 0.8$, $i = 1, \dots, 10$, for Scenario I and II, $a = 3$, $b = 1$, $d = 2$, and $\delta = 0.5$.

Feature	Scenario I				Scenario II			
	Dist.	$h(\cdot)$	$MI(\cdot, C_{0.2})$	$MI(\cdot, C_{0.8})$	Dist.	$h(\cdot)$	$MI(\cdot, C_{0.2})$	$MI(\cdot, C_{0.8})$
V_1	$\text{Unif}(-\frac{1}{2}, \frac{1}{2})$	0	0.5932	0.2932	$\mathcal{N}(0, 1)$	1.4189	0.5520	0.2495
V_2	$\text{Unif}(-\frac{1}{2}, \frac{5}{2})$	1.0986	0.5932	0.2932	$\mathcal{N}(1, 3^2)$	2.5176	0.5520	0.2495
V_3	Y^2	-1.6932	0	0	$\chi_{(1)}^2$	0.7838	0	0
V_4	$\text{Tri}(-1, 1, 0)$	0.5000	0.1785	0.0201	$\mathcal{N}(0, 2)$	1.7655	0.0947	0.0032
V_5	$\text{Unif}(-\frac{1}{2}, \frac{1}{2})$	0	0	0	$\mathcal{N}(0, 1)$	1.4189	0	0
V_6	$Z^2 \stackrel{d}{=} V_3$	-1.6932	0	0	$\chi_{(1)}^2$	0.7838	0	0
V_7	$\text{Unif}(-\frac{1}{2}, \frac{1}{2})$	0	0.0067	0.1153	$\mathcal{N}(0, 1)$	1.4189	0.0124	0.1434
V_8	$X^2 \stackrel{d}{=} V_3$	-1.6932	0	0	$\chi_{(1)}^2$	0.7838	0	0
V_9	$\text{Unif}(\frac{3}{2}, \frac{5}{2})$	0	0	0	$\mathcal{N}(2, 1)$	1.4189	0	0
V_{10}	$\text{Tri}(-1, 1, 0)$	0.5000	0	0	$\mathcal{N}(0, 2)$	1.7655	0	0

From Tables 9 and 10 it is clear that, in many cases, it was not possible to determine an ordering for all ten features. There were only three methods for which this was always possible: MIFS ($\beta \neq 0$), mRMR, and maxMIFS.

These methods achieved the same ordering in all situations: X , Y , Z , $W + 2$, $X - Y$, $Z + W$, $3X + 1$, Y^2 , Z^2 , and X^2 . To understand why, consider Scenario I and the MIFS method (with $\beta = 1$), which is probably the easiest to follow.

- X is selected first because, together with $3X + 1$ has the highest MI with the class, both for $k = 0.2$ and $k = 0.8$. X is selected before $3X + 1$ simply because it was placed first in the list of initial features. As a result, $3X + 1$ becomes redundant and will only be selected in seventh place.
- Y is selected in second place, because it is informative about the class and is not redundant with X . One may think that $X - Y$ should be selected at this step, because $(X, X - Y)$ is a relevance-optimal set and its MI with the class is larger than that of Y , especially when $k = 0.2$. However, $X - Y$ is also more redundant with X . Indeed, for candidate Y , the objective function value is simply $MI(C_{0.2}, Y) = 0.0067$ while, for $X - Y$, is -0.3215, since $MI(C_{0.2}, X - Y) = 0.1785$ and $MI(X, X - Y) = 0.5$.

TABLE 9. Scenario I - Feature selection ordering, (a) $k = 0.2$ and (b) $k = 0.8$. The methods for which the two first selected features form a relevance-optimal set are shown in bold type.

Methods	Order of feature selection									
MIFS ($\beta = 0$)	X	$X-Y$	Y	Z	$W+2$	$Z+W$	-	-	-	-
MIFS ($\beta = .4, .7, 1$)	X	Y	Z	$W+2$	$X-Y$	$Z+W$	$3X+1$	Y^2	Z^2	X^2
MIFS-U ($\beta = 0$)	X	-	-	-	-	-	-	-	-	-
MIFS-U ($\beta = .4, .7, 1$)	X	$3X+1$	$X-Y$	X^2	-	-	-	-	-	-
mRMR	X	Y	Z	$W+2$	$X-Y$	$Z+W$	$3X+1$	Y^2	Z^2	X^2
mMIFS-U	X	$3X+1$	$X-Y$	X^2	-	-	-	-	-	-
MICC	X	X^2	$X-Y$	Y^2	-	-	-	-	-	-
QMIFS	X	$3X+1$	-	-	-	-	-	-	-	-
NMIFS	X	X^2	Y^2	Z^2	$X-Y$	-	-	-	-	-
maxMIFS	X	Y	Z	$W+2$	$X-Y$	$Z+W$	$3X+1$	Y^2	Z^2	X^2
(a) $k = 0.2$.										
Methods	Order of feature selection									
MIFS ($\beta = 0$)	X	Y	$X-Y$	Z	$W+2$	$Z+W$	-	-	-	-
MIFS ($\beta = .4, .7, 1$)	X	Y	Z	$W+2$	$X-Y$	$Z+W$	$3X+1$	Y^2	Z^2	X^2
MIFS-U ($\beta = 0$)	X	-	-	-	-	-	-	-	-	-
MIFS-U ($\beta = .4, .7, 1$)	X	$3X+1$	$X-Y$	X^2	-	-	-	-	-	-
mRMR	X	Y	Z	$W+2$	$X-Y$	$Z+W$	$3X+1$	Y^2	Z^2	X^2
mMIFS-U	X	$3X+1$	$X-Y$	X^2	-	-	-	-	-	-
MICC	X	X^2	$X-Y$	Y^2	-	-	-	-	-	-
QMIFS	X	$3X+1$	-	-	-	-	-	-	-	-
NMIFS	X	X^2	$3X+1$	Z^2	$X-Y$	-	-	-	-	-
maxMIFS	X	Y	Z	$W+2$	$X-Y$	$Z+W$	$3X+1$	Y^2	Z^2	X^2
(b) $k = 0.8$.										

- The next two features to be selected are Z and $W + 2$, both with a null objective function. Note that, after selecting Y , the redundancy of $X - Y$ increases to $\text{MI}(X, X - Y) + \text{MI}(Y, X - Y) = 1$, reinforcing the negative value of its objective function. While selecting Z , features Z^2 , $W + 2$, and $Z + W$ have the same objective function value; they are left behind just because they are placed after Z in the list of initial features.
- At the fifth step, the competition is between $X - Y$ and $Z + W$, because the remaining features, $3X + 1$, Y^2 , Z^2 , and X^2 , all have a $-\infty$ objective function value, since they are fully associated with at least one of the already selected features. $X - Y$ is selected in fifth place and $Z + W$ in sixth, because the former has some association with the class, the latter has not, and they both have the same redundancy.
- Finally, the last selected features, from seventh to tenth place, are $3X + 1$, Y^2 , Z^2 , and X^2 .

As expected, the best methods select first one relevance-optimal set, indeed one of the sets involving independent features. (X, Y) was the chosen one but, depending on the order of features in the initial list, $(3X + 1, Y)$ could also have been selected first. It is worth noting that the next two selected features, Z and $W + 2$, are irrelevant. We argue that this is a good characteristic of the methods given that, in practice, the number of optimal features is unknown and some features beyond the relevant ones may be selected. In general, it is less harmful for the classifier to select irrelevant features. For example, many classifiers require the inversion of the covariance matrix; and while the covariance matrix of (X, Y, Z) is invertible, the one of $(X, Y, X - Y)$ is not.

The MIFS ($\beta \neq 0$), mRMR, and maxMIFS feature selection methods do not suffer from any kind of indeterminations, nor from the possibility of having negative entropies. This cannot be said about the other methods. In the following we give some examples.

TABLE 10. Scenario II - Feature selection ordering, (a) $k = 0.2$ and (b) $k = 0.8$. The methods for which the two first selected features form a relevance-optimal set are shown in bold type.

Methods	Order of feature selection									
MIFS ($\beta = 0$)	X	X-Y	Y	Z	W+2	Z+W	-	-	-	-
MIFS ($\beta = .4, .7, 1$)	X	Y	Z	W+2	X-Y	Z+W	3X+1	Y²	Z²	X²
MIFS-U ($\beta = 0$)	X	X-Y	Y	Z	W+2	Z+W	-	-	-	-
MIFS-U ($\beta = .4$)	X	X-Y	Y	Z	W+2	Z+W	3X+1	Y²	X²	-
MIFS-U ($\beta = .7, 1$)	X	Y	Z	W+2	Z+W	X-Y	3X+1	Y²	X²	-
mRMR	X	Y	Z	W+2	X-Y	Z+W	3X+1	Y²	Z²	X²
mMIFS-U	X	Y	Z	W+2	Z+W	X-Y	3X+1	Y²	X²	-
MICC	X	Y	X-Y	Y²	X²	3X+1	-	-	-	-
QMIFS	X	Y	Z	W+2	Z+W	X-Y	-	-	-	-
NMIFS	X	Y	Z	W+2	X-Y	Z+W	3X+1	Y²	Z²	X²
maxMIFS	X	Y	Z	W+2	X-Y	Z+W	3X+1	Y²	Z²	X²
(a) $k = 0.2$.										
Methods	Order of feature selection									
MIFS ($\beta = 0$)	X	Y	X-Y	Z	W+2	Z+W	-	-	-	-
MIFS ($\beta = .4, .7, 1$)	X	Y	Z	W+2	X-Y	Z+W	3X+1	Y²	Z²	X²
MIFS-U ($\beta = 0$)	X	Y	X-Y	Z	W+2	Z+W	-	-	-	-
MIFS-U ($\beta = .4, .7, 1$)	X	Y	Z	W+2	Z+W	X-Y	3X+1	Y²	X²	-
mRMR	X	Y	Z	W+2	X-Y	Z+W	3X+1	Y²	Z²	X²
mMIFS-U	X	Y	Z	W+2	Z+W	X-Y	3X+1	Y²	X²	-
MICC	X	Y	X-Y	Y²	X²	3X+1	-	-	-	-
QMIFS	X	Y	Z	W+2	Z+W	X-Y	-	-	-	-
NMIFS	X	Y	Z	W+2	X-Y	Z+W	3X+1	Y²	Z²	X²
maxMIFS	X	Y	Z	W+2	X-Y	Z+W	3X+1	Y²	Z²	X²
(b) $k = 0.8$.										

Negative entropies can affect several of the proposed methods. When using NMIFS in Scenario I, both with $k = 0.2$ and $k = 0.8$, X^2 is selected in second place, because the design of the objective function ignored the possibility of a negative entropy. In this case, since the entropy of X^2 is negative, $NI(X^2, X)$ is also negative, which leads to a redundancy contribution that increases the objective function value, something certainly opposite to what the authors have wished. Another problem with negative entropies is that they can lead to indeterminations of the type $+\infty - \infty$. The methods that can be impaired by negative entropies are MIFS-U, mMIFS-U, MICC, QMIFS, and NMIFS.

MIFS ($\beta = 0$) and MIFS-U ($\beta = 0$) always select the features according to the MI between the features and the class, because there is no redundancy component involved in the objective function. For example, in the case of $k = 0.2$ (both scenarios), the order is X , $3X + 1$, $X - Y$, Y , Y^2 , Z , Z^2 , X^2 , $W + 2$, and $Z + W$, according to the values of Table 8. However, if the implementation of the algorithm does not ignore the redundancy component and considers β explicitly, many of the features cannot be selected due to indeterminations of the type $0 \times (+\infty)$. In the cases of MIFS ($\beta = 0$) (both scenarios) and MIFS-U ($\beta = 0$) (Scenario II), only six features can be selected. In Scenario I, MIFS-U behaves even worse. Because the objective function includes β in the numerator and the entropy of an already-selected feature, $h(V_s)$, in the denominator, and since the first selected feature, X , has null entropy, there will always be an indetermination of type $0/0$, and no additional feature can be selected. This shows that special care must be exercised when using MIFS and MIFS-U with $\beta = 0$.

MIFS-U ($\beta = 0.4, 0.7, 1$) and mMIFS-U select as second, third, and fourth features, $3X + 1$, $X - Y$, and X^2 , and from then on are unable to select other features, in Scenario I. The features $3X + 1$, $X - Y$, and X^2

are all selected with a $-\infty$ objective function value; they appear in this order because it is how they appear in the initial list of features. The remaining selection steps are impaired by 0/0 indeterminations. This is due to the quotient $MI(V_i, V_s)/h(V_s)$: since the entropy of X is null, this quotient is either an indetermination, for independent features, or $+\infty$, for fully associated features. Thus, the algorithm selects only features fully associated with X , resulting in a quite wrong selection. As in the case of negative entropies, the methods that are impaired by 0/0 indetermination are MIFS-U, mMIFS-U, MICC, QMIFS, and NMIFS.

The results obtained with MICC and QMIFS in Scenario I are affected by several types of indeterminations simultaneously. For example, QMIFS selects the second feature in the same way as MIFS-U ($\beta = 0.4, 0.7, 1$) and mMIFS-U, i.e., with a $-\infty$ objective function value. However, when trying to select the third feature, the objective function for candidate Y^2 includes 0/0 and $0 \times \infty$ indeterminations, for candidate $X - Y$ includes $+\infty - \infty$, and so on.

When comparing Tables 9 and 10, it is clear that there are fewer problems in Scenario II, i.e., when the random variables have a normal distribution. This is due to the fact that, unlike Scenario I, in Scenario II there are no features with null or negative entropies (Table 8). However, this is not a general property. For this set of features it would be possible to find negative entropies, if $\sigma^2 < (2\pi e)^{-1}$, or null entropies, if $\sigma^2 = (2\pi e)^{-1}$. Despite the improved behavior, in Scenario II it is not possible to terminate the selection process in MIFS ($\beta = 0$), MIFS-U, mMIFS-U, MICC, and QMIFS methods.

The results of MIFS-U, in Scenario II with $k = 0.2$, illustrate an interesting issue regarding the weight given to the redundancy component. With $\beta = 0.4$, $X - Y$ is selected in second place and Y in third, while with $\beta = 0.7$ and $\beta = 1$ the opposite occurs. Indeed, the association with the class is higher for $X - Y$ than for Y . But Y has no redundancy with X while $X - Y$ has some. For $\beta = 0.4$, the redundancy component is given relatively low weight, not enough for cancelling out the strength of the relevancy component, and $X - Y$ gets selected before Y .

A general conclusion can be drawn from this study. From the analysis of the objective functions and from the results of Tables 9 and 10, there are only three methods that do not have problems with indeterminations, which are MIFS ($\beta \neq 0$), mRMR, and maxMIFS. However, due to the problems with MIFS ($\beta = 0$), we cannot recommend its use with small β values. In all these three methods, the same order was obtained in all cases. The WR independent features, X and Y are selected first, then the irrelevant ones, Z and $W + 2$, and finally features that have become redundant with already-selected ones.

8. SIMULATION STUDY

To assess the performance of the eight feature selection methods (MIFS, MIFS-U, mRMR, mMIFS-U, QMIFS, MICC, NMIFS, and maxMIFS), as well as the importance of correctly estimating the MI, a simulation study based on the two evaluation scenarios presented in Section 5 was developed. We randomly generated 5000 samples of sizes $n = 50, 100, 500, 1000$, and 5000, for the ten input features shown in Table 3, and applied the eight feature selection methods to each sample.

In Table 11 we compare our MI estimates with those obtained in [21] and [25], and with the true value obtained with the results of Section 6. The case considered in [21] and [25] was that of Scenario I with $k = 0.2$, and a sample size $n = 1000$. Our MI estimates were obtained indirectly using property (e) of Section 2, for $\hat{MI}(C_{0.2}, V_i)$, and equation (d) of the same section, for $\hat{MI}(V_j, V_i)$. The differential entropies were estimated by partitioning the simulated values in equal-width bins. We added the correction factor $\ln(\Delta)$ to the entropies estimated based on the discretized values, where Δ is the length of the bins (vide [19, 29] for details). The number of bins considered is a function of the sample size given by $m = \lceil \sqrt{n} \rceil$. As in [21] and [25], we present the mean of the MI estimates between the class and the first four features, $\hat{MI}(C_{0.2}, V_i)$, $i = 1, 2, 3, 4$, the mean of the MI estimates between the first feature, and the third and fourth features, $\hat{MI}(V_1, V_i)$, $i = 3, 4$. The true values were extracted from Table 8. It is clear from Table 11 that the results of [21] and [25] have large deviations relative to the true values. Our own estimates are much closer. We guess that these errors may be attributed to the fact that [21] and [25] only considered one sample, which is clearly insufficient

for statistical confidence. Another possibility is that they either have used the same bin width in the discretization of univariate and bivariate distributions, or have not included the corresponding correction factor. We believe that estimation errors could be one reason explaining the quality overstatement of some proposed methods. In [34], we have extended this study to other entropy and MI estimates for which we have estimated the mean square error, $\text{MSE}(\hat{h}(V_i))$, $i = 1, \dots, 8$, and $\text{MSE}(\hat{\text{MI}}(C_k, V_i))$, $k = 0.2, 0.8$, $i = 1, \dots, 8$, for scenarios I and II. In all cases the estimated values are close to the true ones.

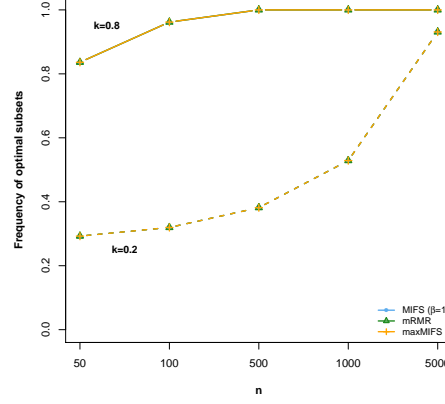
TABLE 11. Estimated (mean of the) MI between the class, $C_{0.2}$, and the first four features, V_1, V_2, V_3 and V_4 , $\hat{\text{MI}}(C_{0.2}, V_i)$, $i = 1, 2, 3, 4$, and (mean of the) MI between the first feature, V_1 , and the third and fourth features, V_3 and V_4 , $\hat{\text{MI}}(V_1, V_i)$, $i = 3, 4$, considering sample(s) of size $n = 1000$.

Source	$\hat{\text{MI}}(C_{0.2}, V_i), i = 1, 2$	$\hat{\text{MI}}(C_{0.2}, V_3)$	$\hat{\text{MI}}(C_{0.2}, V_4)$	$\hat{\text{MI}}(V_1, V_3)$	$\hat{\text{MI}}(V_1, V_4)$
[21]	0.8459	0.0170	0.2621	0.0610	0.6168
[25]	0.8438	0.0383	0.2807	0.0634	0.6099
Our estimate	0.5932	0.0075	0.1779	0.0107	0.5004
True	0.5932	0	0.1785	0	0.5

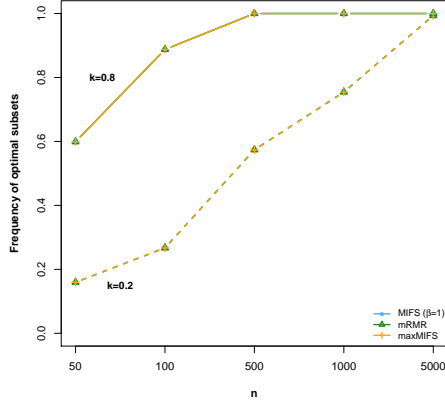
TABLE 12. Relative frequency of the optimal subsets.

Methods	Scenario I		Scenario II	
	$k = 0.2$	$k = 0.8$	$k = 0.2$	$k = 0.8$
MIFS ($\beta = 0$)	0	0	0	0
MIFS ($\beta = 0.4$)	0.9744	1	0.9998	1
MIFS ($\beta = 0.7$)	0.9502	1	0.9982	1
MIFS ($\beta = 1$)	0.9310	1	0.9936	1
MIFS-U ($\beta = 0$)	0	0	0	0
MIFS-U ($\beta = 0.4$)	0	0.4024	0	0.6720
MIFS-U ($\beta = 0.7$)	0.3950	0.4024	0.6556	1
MIFS-U ($\beta = 1$)	0.3964	0.4024	0.6556	1
mRMR	0.9310	1	0.9936	1
mMIFS-U	0.3964	0.4024	0.6736	1
MICC	0	0.0140	0.8582	0.9844
QMIFS	0.3964	0.4024	0.6736	1
NMIFS	0	0	0.2638	1
maxMIFS	0.9310	1	0.9936	1

In Figures 1(a) and 1(b) we study the performance of the three best feature selection methods, MIFS ($\beta = 1$), mRMR, and maxMIFS, as function of the sample size, for both scenarios and using $k = 0.2$ and $k = 0.8$. The performance is evaluated through the estimated probability of selecting first any of the relevance-optimal sets, i.e., (X, Y) , $(X, X - Y)$, $(Y, X - Y)$, $(3X + 1, Y)$, or $(3X + 1, X - Y)$. The results show that the performance can be significantly dependent on the sample size, especially for $k = 0.2$. For $k = 0.8$ the probability is close to one for n equal or higher than 500, meaning that in most cases the features are well selected. However, for $k = 0.2$, good results are only obtained when n is 5000. This is due to the lack of precision in estimating the entropy and the MI. Indeed, the case $k = 0.2$ is the most challenging one. Given that the strength of Y is smaller in the class definition, the MI between the class and Y is also smaller. For example, in Scenario I the (theoretical) MI is 0.0067 for $k = 0.2$ and 0.1153 for $k = 0.8$, as shown in Table 8. Thus, the possibility of selecting in second place features other than Y is much higher. Consequently, errors in estimating the entropy and the MI, and their propagation in the calculation of the objective function, which occur with smaller sample sizes, lead to smaller performance.



(a) Scenario I



(b) Scenario II

FIGURE 1. Estimated probability of selecting a relevance-optimal pair first, for MIFS ($\beta = 1$), mRMR, and maxMIFS, considering $k = 0.2$ and $k = 0.8$, in (a) Scenario I and (b) Scenario II.

In Table 12 we compare the performance of the 8 methods for the largest sample size, $n = 5000$. These results confirm the poor performance of MIFS ($\beta = 0$), MIFS-U, mMIFS-U, MICC, QMIFS, and NMIFS, especially in Scenario I. There are even some cases where it was never possible to select a relevance-optimal pair of features first. This is due to unstable numerical behavior around the indeterminations of their objective functions. A probability of 0 was expected for MIFS and MIFS-U with $\beta = 0$. In these cases, $\beta = 0$ removes the redundancy component, and the first two selected features are always X and $3X + 1$, the ones mostly associated with the class. But NMIFS also was never able to select correctly the first two features in Scenario I. In this case, the second selected feature was always $3X + 1$, because the entropy estimate of X , despite being close to zero, was negative in all samples. Thus, the redundancy component is always negative, forcing the objective function to be maximized by the candidate feature that is most associated with already selected features, which is $3X + 1$. This case illustrates the problems motivated by the sensitivity of objective functions.

9. CONCLUSIONS

This paper undergoes an evaluation of the feature selection methods based on MI and two-dimensional sequential forward search, i.e. MIFS, MIFS-U, mRMR, mMIF-U, MICC, QMIFS, NMIFS, and maxMIFS. We clarified the differences between the MI and entropy properties of discrete and continuous random variables, whose misunderstanding has been a source of error in the proposal of feature selection methods. To support our evaluation, we established clear definitions for the notions of relevant, redundant, and irrelevant features, an issue still displaying controversy in the literature. Then, we developed a theoretical framework that allows obtaining the true feature ordering for each method under analysis, based on a scenario with two classes and a carefully chosen representative set of features. The ordering obtained in this way does not depend on entropy or MI estimation methods, classifiers, or specific datasets, leading to an undoubtful comparison of the methods, which is the main contribution of our work. The feature ordering was also compared with the optimal subsets of features, obtained using the notion of relevance-optimal sets. Moreover, our framework unveiled inconsistencies in the construction of the objective functions of several feature selection methods, due to various types of indeterminations and the possibility of entropies taking null or negative values in the case of continuous random variables.

10. ACKNOWLEDGEMENTS

This work was partially funded by Fundação para a Ciência e a Tecnologia (FCT) through project UID/Multi/ 04621/2013. Cláudia Pascoal also acknowledges the support of FCT via PhD grant SFRH/BD/42547/2007.

REFERENCES

1. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res* ; .
2. Ferri F, Pudil P, Hatef M, Kittler J. Comparative study of techniques for large scale feature selection. *Pattern Recognition in Practice IV*, Gelsema E, Kanal L (eds.). Elsevier Science B.V., 1994; 403–413.
3. Dash M, Liu H. Feature selection for classification. *Intell Data Anal* 1997; **1**(3):131–156.
4. Jain A, Zongker D. Feature selection: Evaluation, application, and small sample performance. *IEEE T Pattern Anal* 1997; **19**:153–158.
5. Kudo M, Sklansky J. Comparison of algorithms that select features for pattern classifiers. *Pattern Recogn* January 2000; **33**(1):25–41.
6. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007; **23**(19):2507–2517.
7. Brown G, Pocock A, Zhao M, Luján M. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *J Mach Learn Res* March 2012; **13**:27–66.
8. Vergara JR, Estévez PA. A review of feature selection methods based on mutual information. *Neural Comput Appl* 2014; **24**(1):175–186.
9. Bennasar M, Hicks Y, Setchi R. Feature selection using joint mutual information maximisation. *Expert Syst Appl* 2015; **42**(22):8520–8532.
10. Freeman C, Kulić D, Basir O. An evaluation of classifier-specific filter measure performance for feature selection. *Pattern Recogn* 2015; **48**(5):1812–1826.
11. Vinh NX, Zhou S, Chan J, Bailey J. Can high-order dependencies improve mutual information based feature selection? *Pattern Recogn* 2016; **53**:46–58.
12. John G, Kohavi R, Pfleger K. Irrelevant features and the subset selection problem. *Proceedings of the 11th International Conference in Machine Learning, Rutgers University in New Brunswick, NJ, USA, 10-15 July*, Morgan Kaufmann, 1994; 121–129.
13. Koller D, Sahami M. Toward Optimal Feature Selection. *Proceedings of the 13th International Conference on Machine Learning, ICML'96, Bari, Italy, 3-6 July*, Saitta L (ed.), Morgan Kaufmann Publishers, 1996; 284–292.

14. Blum A, Langley P. Selection of Relevant Features and Examples in Machine Learning. *Artif Intell* December 1997; **97**(1-2):245–271.
15. Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy. *J Mach Learn Res* December 2004; **5**:1205–1224.
16. Kohavi R, John G. Wrappers for feature subset selection. *Artif Intell* 1997; **97**(1-2):273–324.
17. Sindhwani V, Rakshit S, Deodhar D, Erdogmus D, Príncipe J, Niyogi P. Feature selection in mlps and svms based on maximum output information. *IEEE T Neural Networ* July 2004; **15**(4):937–948.
18. Lal TN, Chapelle O, Weston J, Elisseeff A. *Embedded methods*. Studies in Fuzziness and Soft Computing; 207, Springer: Berlin, Germany, 2006; 137–165.
19. Cover T, Thomas J. *Elements of Information Theory*. 2nd edn., Wiley Sons: New York, NY, USA, 2006.
20. Battiti R. Using mutual information for selecting features in supervised neural net learning. *IEEE T Neural Networ* July 1994; **5**(4):537–550.
21. Kwak N, Choi C. Input feature selection for classification problems. *IEEE T Neural Networ* January 2002; **13**(1):143–159.
22. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE T Pattern Anal* 2005; **27**:1226–1238.
23. Estévez P, Tesmer M, Perez C, Zurada J. Normalized mutual information feature selection. *IEEE T Neural Networ* February 2009; **20**(2):189–201.
24. Novovicová J, Somol P, Haindl M, Pudil P. Conditional mutual information based feature selection for classification task. *Proceedings of the 12th Iberoamerican Conference on Progress in Pattern Recognition Image Analysis and Applications, CIARP'07, Viña del Mar/Valparaiso, Chile, 13-16 November, Lecture Notes in Computer Science*, vol. 4756, Rueda L, Mery D, Kittler J (eds.), Springer-Verlag: Berlin, Heidelberg, 2007; 417–426.
25. Huang JJ, Lv N, Li SQ, Cai YZ. Feature selection for classificatory analysis based on information-theoretic criteria. *Acta Automatica Sinica* 2008; **34**(3):383–392.
26. Kotz S. Recent Results in Information Theory June 1966; **3**(1):1–93.
27. Brillinger D. Some data analyses using mutual information. *Braz J Probab Stat* 2004; **18**(6):163–183.
28. Tsujishita T. On triple mutual information. *Adv Appl Math* 1995; **16**(3):269–274.
29. Pascoal C. Contributions to Variable Selection and Robust Anomaly Detection in Telecommunications. PhD Thesis, Instituto Superior Técnico, Technical University of Lisbon 2014.
30. Lazo A, Rathie P. On the entropy of continuous probability distributons. *IEEE T Inform Theory* 1978; **IT-24**(1):1–93.
31. Azzalini A. A class of distributions which includes the normal ones. *Scand J Stat* 1985; **12**(2):171–178.
32. Dadkhah K, Midi H, Olimjon S. The performance of mutual information for mixture of bivariate normal distributions based on robust kernel estimation. *Appl Math Sci* 2010; **4**(29):1417–1436.
33. Wolfram S. *The Mathematica book*. 5th edn., Wolfram Media: New York, 2003.
34. Pascoal C, Oliveira MR, Valadas R, Filzmoser P, Salvador P, Pacheco A. Robust feature selection and robust PCA for Internet traffic anomaly detection. *Proceedings of the 31st IEEE International Conference on Computer Communications, INFOCOM'12, Orlando, FL, USA, 25-30 March*, Greenberg A, Sohrawy K (eds.), IEEE, 2012; 1755–1763.